ETL Specification Document

Project Name: Customer 360 Analytics Platform

Document Version: 1.0

Prepared By: Data Engineering Team

Date: May 12, 2025

# Medallion Architecture Overview

|  |  |  |
| --- | --- | --- |
| Layer | Purpose | Data Granularity |
| Bronze | Raw ingestion of source data with minimal transformation | Raw, uncleaned data |
| Silver | Cleaned, normalized, and validated data | Business entity level |
| Gold | Aggregated, business-ready data for analytics | Summary/aggregate level |

# Source Systems

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source Name | Source Type | Table Name | Description | Frequency |
| CRM\_DB | MySQL | customer\_master | Contains customer profile data | Daily |
| TRANSACTION\_DB | PostgreSQL | transactions | Customer purchase transactions | Hourly |
| WEB\_LOGS | CSV on S3 | web\_clickstream | Clickstream logs from the website | Daily |

## Table: bronze\_customer\_master

|  |  |  |
| --- | --- | --- |
| Field Name | Data Type | Description |
| customer\_id | INT | Unique customer ID |
| full\_name | STRING | Customer's full name |
| email | STRING | Email address |
| signup\_date | DATE | Date of account creation |

## Table: bronze\_transactions

|  |  |  |
| --- | --- | --- |
| Field Name | Data Type | Description |
| transaction\_id | INT | Unique transaction ID |
| customer\_id | INT | FK to customer table |
| product\_id | INT | ID of the purchased product |
| amount | DECIMAL | Transaction amount |
| txn\_timestamp | TIMESTAMP | Transaction time |

## Table: silver\_customer

|  |  |  |
| --- | --- | --- |
| Field Name | Data Type | Transformation Logic |
| customer\_id | INT | Direct mapping from bronze |
| first\_name | STRING | SPLIT(full\_name, ' ')[0] |
| last\_name | STRING | SPLIT(full\_name, ' ')[1] |
| email | STRING | Lowercase and validated |
| signup\_year | INT | YEAR(signup\_date) |
| is\_email\_valid | BOOLEAN | email LIKE '%@%.%' |

## Table: silver\_transaction

|  |  |  |
| --- | --- | --- |
| Field Name | Data Type | Transformation Logic |
| transaction\_id | INT | Direct mapping from bronze |
| customer\_id | INT | Direct mapping |
| amount | DECIMAL | Null-safe cast to decimal |
| txn\_date | DATE | CAST(txn\_timestamp AS DATE) |
| txn\_hour | INT | HOUR(txn\_timestamp) |

## Table: gold\_customer\_summary

|  |  |  |
| --- | --- | --- |
| Field Name | Data Type | Transformation Logic |
| customer\_id | INT | Group by customer |
| total\_spent | DECIMAL | SUM(amount) |
| total\_transactions | INT | COUNT(transaction\_id) |
| avg\_transaction\_amount | DECIMAL | AVG(amount) |
| first\_transaction\_date | DATE | MIN(txn\_date) |
| last\_transaction\_date | DATE | MAX(txn\_date) |

## Transformation Mapping Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source Table | Source Field | Transformation Logic | Target Table | Target Field |
| bronze\_customer\_master | full\_name | SPLIT(full\_name, ' ')[0] | silver\_customer | first\_name |
| bronze\_customer\_master | full\_name | SPLIT(full\_name, ' ')[1] | silver\_customer | last\_name |
| bronze\_transactions | txn\_timestamp | CAST(txn\_timestamp AS DATE) | silver\_transaction | txn\_date |
| silver\_transaction | amount | SUM(amount) GROUP BY customer\_id | gold\_customer\_summary | total\_spent |
| silver\_transaction | transaction\_id | COUNT(transaction\_id) | gold\_customer\_summary | total\_transactions |
| silver\_transaction | amount | AVG(amount) | gold\_customer\_summary | avg\_transaction\_amount |

## Schedule and Frequency

|  |  |  |
| --- | --- | --- |
| ETL Layer | Schedule | Tool/Framework |
| Bronze | Hourly | Apache Spark |
| Silver | Hourly | Apache Spark |
| Gold | Daily @ 12 AM | DBT or Airflow |

## Data Quality Checks

|  |  |  |
| --- | --- | --- |
| Check Type | Target Table | Rule |
| Null check | silver\_customer | customer\_id IS NOT NULL |
| Email format | silver\_customer | email LIKE '%@%.%' |
| Duplicate detection | silver\_transaction | transaction\_id NOT IN (SELECT ...) |
| Amount > 0 | silver\_transaction | amount > 0 |